PCCP

PAPER

Cite this: *Phys. Chem. Chem. Phys.*, 2019, **21**, 6544

Design of a structure-based model for protein folding from flexible conformations†

Ana M. Rubio and Antonio Rey *

Received 10th January 2019,
Accepted 13th February 2019

DOI: 10.1039/c9cp00168a

rsc.li/pccp

The use of coarse-grained models is important in many fields, especially those that use computer simulation to analyze large systems in processes that span long-time scales, as happens in protein folding. Among those approaches, structure-based models have been widely and successfully used for a few decades now. They usually take a single native conformation, experimentally solved, of the protein studied to determine the native contacts, which subsequently define the interaction potential for the simulation. The characteristics of the folding transition can then be analyzed from the computed trajectories. In this paper, we consider the possibility of enriching these models by considering the structural fluctuations present in the native state of a globular protein at room temperature in an aqueous environment. We use the different conformers experimentally provided when the protein structure was determined by nuclear magnetic resonance (NMR) spectroscopy as an approximate ensemble to test our methodology, which includes the definition of a global interaction potential and the analysis of the thermodynamic and structural characteristics of the folding process. The results are compared with traditional, single structure models.

Introduction

Structure-based (or Gō-type) folding models have been used in the last few decades to analyze the characteristics of the protein folding process.[1,2] In combination with coarse-grained models, which reduce the complexity of the protein representation,[3,4] or even with atomistic models,[5] they have provided useful insights in fields including folding thermodynamics, kinetics, pathways, *etc.* On the other hand, since structure-based models largely simplify the protein stabilizing interactions, reducing them to those present in the native state, these models cannot properly reproduce the unfolded state, or the presence of folding intermediates that include non-native interactions.

A structure-based interaction model relies on the comparison between every sampled conformation of the polypeptide chain and the native state. This state is usually considered at the level of a native contact map, *i.e.*, a 2D representation of the native conformation in which two residues in contact in the native state (according to some predefined geometrical criterion) are marked as a spot in the map; see several examples in the figures below. The native state is customarily taken from the experimentally determined protein structures deposited in the Protein Data Bank (PDB).[6] About 90% of the currently available structures correspond to X-ray diffraction data, where the protein has been previously crystallized; thus, a single conformation for the native state is deposited. On the other hand, for nuclear magnetic resonance (NMR) determined native structures, the measurements are made in solution, where certain regions of the native structure may show different degrees of flexibility due to thermal fluctuations. Since the NMR measurement time is long in comparison to the time scale of these fluctuations, the raw data must be processed afterwards, and several conformers (or ''NMR models'', as named in the PDB) are deposited. Of course, they do not need to be an accurate reflection of the flexibility corresponding to the native state in thermodynamic equilibrium with its aqueous environment. The proper interpretation of the NMR spectra for the definition of spatial restraints, their character (strong, medium or weak), and the numerical method used to get the final structure may have an influence on the final output.[7] The NMR models are usually ranked in the PDB file by quality (fewer violations of experimental restraints), so the first NMR model is usually considered in simulations when only one structure is needed. This is the usual strategy when using NMR data of proteins taken from the PDB in structure-based models,[8] or as a matter of fact, also in atomistic simulations with different physical force fields.[9] Thus, the information about the structural flexibility, even in an approximate form provided by the NMR experiment, is lost.

In the last decade, only a few studies have tried to go beyond this situation. In our group, we took advantage of a very simple

Departamento de Química Física, Facultad de Ciencias Químicas, Universidad Complutense de Madrid, E-28040 Madrid, Spain. E-mail: areygayo@ucm.es

† Electronic supplementary information (ESI) available. See DOI: 10.1039/c9cp00168a

coarse-grained strategy to individually consider every NMR model[10] and get what tries to be a more complete picture of the folding process by combining the different outputs. At the same time, we realized that different folding pathways could appear when considering different NMR models, something that does not make any sense since all of them correspond to the very same protein under identical conditions. Jiang and Hansmann[11] proposed to use the conformation in the PDB file that best represents the set of NMR models, which does not necessarily correspond to the first one; some native contacts could also be removed or added to the contact map of this conformation to gather together information from the full ensemble of NMR models. More recently, Lammert *et al.*[12] have tried to analyze the folding of a very flexible protein by using the X-ray structure to define the native contact map, but suppressing from it the contacts in those regions that, according to the NMR experimental data for the same protein, correspond to highly flexible loops or tails in solution. For proteins showing slow conformational transitions, potentials with different minima corresponding to the different states have also been tested.[13–15] Some work considering allosteric transitions has also been done with the same premise.[16]

In this work, we try a different approach, trying to consider the available experimental flexibility information in a more comprehensive way. Therefore, we consider the full native structure as potentially flexible, and tentatively rely on the different NMR conformers to define the degree of conservation (and therefore the strength) of the native contacts in the interaction potential of the simulation model introduced in this work. Even though, as already mentioned, the NMR conformers may not properly represent the equilibrium microstates of the native protein, we can use them as a proof of concept of the model design.

Simulation and interaction model

In this manuscript, we represent our polypeptide chain as a linear polymer of beads centered at the position of the α-carbons of the different residues. They are thus joined by segments of equal length (3.8 Å, which corresponds to a *trans* peptide bond). The conformations are sampled by a Monte Carlo procedure, previously described,[17,18] coupled to a replica exchange method[19] where different replicas are simulated at different temperatures (parallel tempering). The Monte Carlo sampling includes both local (single bead) and collective (multi-bead) movement trials, to generate configurations of the system that correspond to the conformational equilibrium distribution of the protein model at each of the temperatures included in the parallel tempering procedure.

Thus, we focus on thermodynamic and structural properties to study the folding/unfolding process as a function of temperature. In previous work that used a single conformation to define the native contacts, we defined the model interactions as a truncated harmonic well centered at the native distance and whose depth, equal for all the native contacts, defines the energy unit of the model. Its mathematical definition is

$$u_{ij}(r_{ij}) = \begin{cases} \varepsilon\left[\dfrac{\left(r_{ij} - d_{ij}^{\mathrm{nat}}\right)^2}{a^2} - 1\right] & \text{if } d_{ij}^{\mathrm{nat}} - a < r_{ij} < d_{ij}^{\mathrm{nat}} + a \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

In eqn (1), r_{ij} is the distance between the beads representing residues i and j, and d_{ij}^{nat} is the corresponding distance between their α-carbons in the native state; $\varepsilon = 1$ defines the energy unit for the model, and a indicates the width of the attractive well for the native contacts. Values of $a = 0.5$ or 0.6 Å have provided correct results for different proteins in previous work from our group.[20–22]

In addition to the attractive interactions for pairs of residues involved in native contacts, we use an excluded volume term acting on every pair of non-neighbor model beads, to avoid unphysical overlappings in the sampled conformations.[17]

As stated in the Introduction, we have considered an NMR-determined protein structure as an example of a set of flexible conformers compatible with a given protein native state. Following the recent work from Onuchic *et al.*,[12] we have considered S6 ribosomal protein, which appears with the PDB code 2KJV with 20 NMR models deposited by the experimental group.[23] This protein's structure has also been solved by X-ray diffraction (with 4 residues less at the C-terminal end) and is deposited as a single conformer with the PDB code 1RIS.[24] For the sake of comparison, the native structure and contact map for 1RIS are shown as Fig. S1 in the ESI.† The native structure has two α-helices packed against a 4-stranded β-sheet, with the elements of secondary structure distributed along the sequence as β1–α1–β2–β3–α2–β4.

In Fig. 1, we summarize the main characteristics of the 20 NMR conformers in PDB file 2KJV. We show a superposition of the 20 NMR models drawn with VMD,[25] a global quantitative comparison among them, computed as the root mean square deviation, RMSD, between every pair of NMR models after optimal superposition (at the level of α-carbons), and the contact map of the first NMR model. The data show a less compact structure in NMR than in X-ray: the radius of gyration, at the level of α-carbons again, is 12.95 Å in 1RIS, and oscillates between 15.1 and 15.5 Å for the NMR models in 2KJV. In addition, the number of non-local native contacts (with $|i - j| \geq 4$) is 212 in the X-ray data, and between 124 and 145 for the different NMR models. The structural fluctuations are mostly centered in the residues of the C-terminal tail and in the loop situated at the very middle of the chain, roughly between residues 45 and 55. This corresponds to the region between strands β2 and β3, which become shorter in the NMR conformations than in the X-ray ones.

At the level of native contacts, we have tried to visualize these fluctuations by (1) defining the contact matrices of the different NMR models, where a 0 means the absence of native contact and 1 means a native contact between residues i and j (as in previous work, we consider a native contact to appear

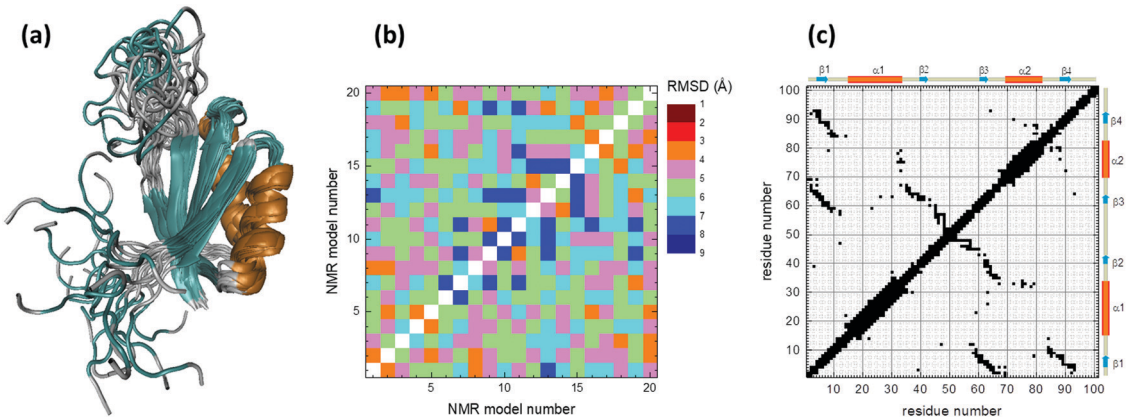


Fig. 1 Structural information of S6 protein in PDB file 2KJV: (a) superposition of the 20 NMR conformers; (b) structural differences among the 20 NMR conformers, computed as RMSD (Å) between α-carbons; (c) native contact map of the first conformer, NMR-01; a sketch representation of the secondary structure elements is included along the axes.

when the distance between any of the heavy – non-hydrogen– atoms belonging to i and j is less than 4.5 Å); and (2) adding up the 20 matrices. The result is plotted in Fig. 2a. We find that the two α-helices in the structure are very well defined in all the NMR models and can therefore be considered as rigid (or less flexible) regions in the native protein. The same happens to many of the contacts between strands β1–β3 and β1–β4 in the β-sheet. On the other hand, the β2–β3 packing shows high variability, as depicted by many of its contacts appearing in less than half of the NMR models. This is especially so for those contacts closer to the loop (to the main diagonal, in Fig. 2a). The same happens to the contacts between helices α1 and α2. The contacts between the helices and the β-sheet seldom occur in the set of NMR models, justifying the less compact native structure already commented on. Finally, the last residues at the C-terminal show no native contacts or, as much, a few sporadic ones present in a single NMR model.

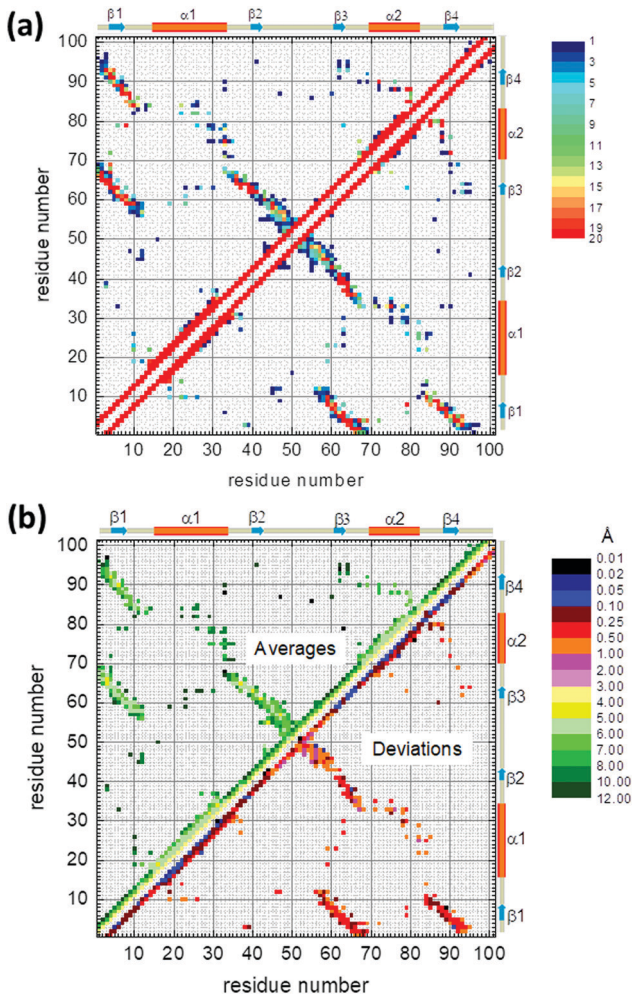


Fig. 2 (a) Contact map for the global NMR model, color-code indicating the number of NMR conformers in which a given native contact appears. (b) Average native distances (upper left triangle) and their statistical deviations (lower right triangle) among the 20 NMR conformers in PDB file 2KJV.

In the standard version of structure-based models, where the protein sequence is disregarded, all the native contacts are given the same energy, as can be seen in eqn (1). In this work, we use the colored contact map of Fig. 2a as the set of interaction energies for the native contacts. Just to keep the energy scale similar to the cases based on a single structure, we have normalized the added contact matrix, so that for the contacts with $|i-j| \geq 4$ appearing in any of the NMR models, we define:

$$\varepsilon_{ij\ \mathrm{contact}} = \frac{\text{number of NMR models with the } ij \text{ contact present}}{\langle \text{number of NMR models for native contacts} \rangle} \quad (2)$$

The average $\langle \cdots \rangle$ in the denominator is computed over the full set of contacts. In the NMR models of PDB file 2KJV, this average equals 8.7 NMR models. Therefore, in our simulation model, the contacts appearing in more rigid regions are considered more stabilizing (larger attractive energy) than those appearing in more flexible regions. Once more, this scale should be physically sound only if the set of conformers chosen to define the flexibility is representative of the equilibrium population of microstates, and that is not necessarily true for the NMR models in the PDB file. However, we consider that file

here to represent a test case of our interaction model. For $|i-j| = 2$ (virtual bond angles) and $|i-j| = 3$ (virtual torsion angles), all the weights are the same, $\varepsilon = 1$. In the latter case, we consider the chirality of the torsion angle to avoid getting the mirror image of the folded state.[17]

As can be seen in eqn (1), the definition of a structure-based potential includes not only the energy term favoring the native state, but also its geometry. This is considered in such a way that the energy for every ij native contact becomes minimum when $r_{ij} = d_{ij}^{\mathrm{nat}}$. Probably, this is the main technical reason why the use of multiple conformations has been previously avoided. In this situation, these d_{ij}^{nat} distances are different in every NMR model, which precludes a proper mathematical definition of the energy terms. In previous works already mentioned, the problem has been either avoided, by using only the distances in the X-ray structure,[12] or simplified using the largest distance among the NMR models for a given contact.[11]

To analyze this fact in our test case, in Fig. 2b, we show also a map where the average distances (upper left triangle) and their statistical deviations (lower right triangle) are shown for the native contacts. They are computed on the α-carbon atoms, the only ones included in the simulation model afterwards. Since, as we have seen, many native contacts appear in just a few NMR models, these statistics are computed over the full set (20 cases for 2KJV) of conformers included in the PDB file. This tries to partially reduce the influence of the specific cut-off distance used to define a native contact, and especially to avoid irrelevant statistics based on a very small number of NMR models for most of the contacts in flexible regions. As can be seen in Fig. 2b, the average distances of the native contacts are in the range of 4–5 Å for α-carbons in neighbor β-strands (where the contacts frequently involve the atoms in the protein backbone) to values as high as 10 Å for regions involving α-helix–α-helix or α-helix–β-sheet contacts, where the real atoms in contact belong to the residue side-chains, and the α-carbons are farther apart. The deviations follow a trend that, as expected, mirrors the flexibility already seen in the NMR models. The smaller values, well beyond 0.5 Å, correspond to rigid sections, while mobile parts create deviations larger than 1 Å.

When a full atom representation of the protein is considered, the flexibility must take into account the real rotamers appearing in the torsional angles of the backbone and side chains, which cannot adopt arbitrary conformations. Plain averaging can then be especially dangerous if the average value corresponds to an impossible conformation.[26] However, in a coarse-grained definition of the protein geometry, as we are considering here, a simpler approach can be taken. Therefore, we center the attractive wells for the native contacts at the average distance values, and adjust their width according to the standard deviations, so that all the native distances in the different NMR models are included in the attractive wells. Specifically, for every native contact between residues i and j (including virtual bond and torsion angles), our model defines the width of eqn (1) as contact dependent,

$$a_{ij}\ (\text{Å}) = 0.25 + \frac{\sigma_{\mathrm{ij}}}{2}, \tag{3}$$

where σ_{ij} represents the deviations shown in Fig. 2b. The additive factor 0.25 has been optimized by checking values between 0.2 and 0.5 Å. Large values result in wider attractive wells that, as we have previously checked, create less cooperative folding/unfolding transitions.[17,27] That is also the reason why we are dividing the deviations by a factor of 2.

With all these considerations, the sampling of the proteins with our Monte Carlo and parallel tempering method is the same as what we have previously described in works considering a single structure.[18] The results presented here correspond to the average of 8 independent runs in every case. In each run, we use 40 to 50 temperatures in the parallel tempering scheme, depending on the complexity of the folding process found for every interaction model considered, to warrant a proper traveling of the replicas along the full set of temperatures. At every temperature, 3×10^8 Monte Carlo steps (conformations) are sampled for thermalization, and 10^9 additional Monte Carlo steps are computed for data recording.

Results

S6 is a protein that, according to the available experimental evidence, folds in a two-state process from a thermodynamic point of view, with a relatively high free energy barrier between the native and the unfolded state.[28] The folding pathway, on the other hand, is rather complex, and it has been the focus of extensive experimental and simulation work.[29–31] Actually, S6 protein and its ''circular permutants''[32] have been the subject of a large deal of protein folding research, even using some models that try to account for its flexibility.[12] In this work, we will focus on the wild-type protein alone, since the conformational richness available in the PDB file 2KJV, which we are taking as an approximation of the flexibility of the native structure in solution, could be for sure affected by any sequence engineering transformation. For this reason, we mainly focus our presentation of results on those related to the flexibility of the folded state and not so much on the proper reproduction of all the experimental evidence for this protein, but being certain that the main experimental features of the folding transition are preserved.

For comparison with our new methodology, we also have run our standard model based on a single native structure for S6, both from the X-ray data and from every one of the 20 NMR models, considered individually. In Fig. S2 of the ESI†, we show the heat capacity curves as a function of temperature (both in reduced units) we have obtained for all these cases. The results from the single X-ray structure (dashed black curve) are consistent with a thermodynamic two-state transition: a high and narrow peak at the transition temperature (T_{m}) between the folded and the unfolded states. The results from the individual NMR conformers, on the other hand, show a wide spectrum of possibilities, including wide folding/unfolding peaks and, in some cases, a second peak at lower temperatures. The transition temperature for the X-ray structure is significantly larger than those obtained from the NMR conformers,

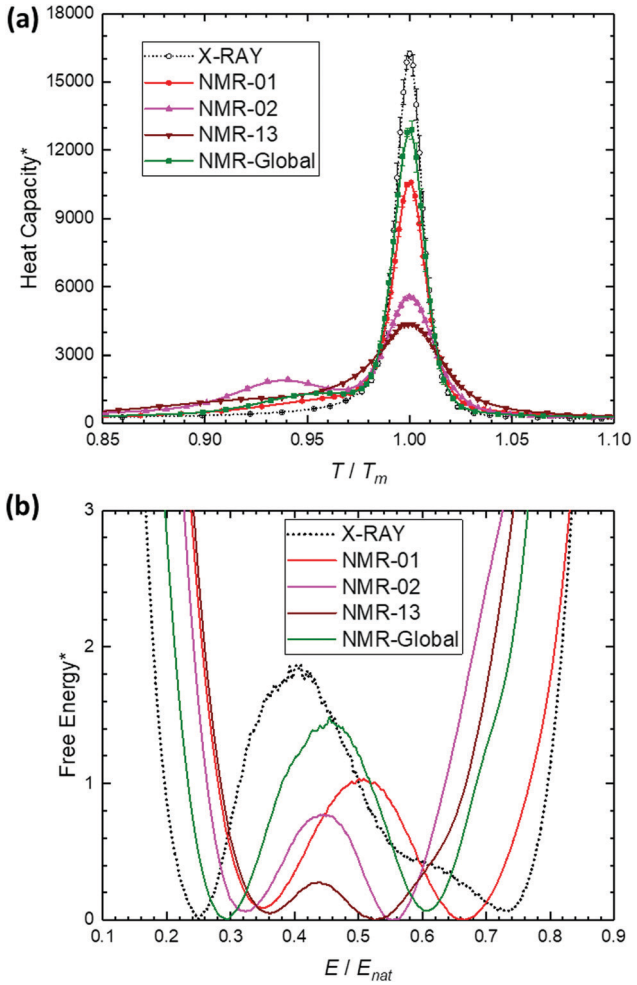


Fig. 3 (a) Reduced heat capacity as a function of the reduced temperature resulting for the different simulation models as indicated in the legend. (b) Reduced free energies at the transition temperature, as a function of the system energy, normalized by the native energy, for the different simulation models.

mainly reflecting the larger number of native contacts in the former, as we have already mentioned, which are lost upon unfolding. To properly compare among the different cases, in Fig. 3a, we show these heat capacity curves against the reduced temperature T/T_m, where T_m corresponds to every individual case. To make this analysis clearer, we have selected a few of the NMR cases: model 01, the one usually considered in standard structure-based models; model 02, which shows the least average RMSD with the rest of the NMR models (see Fig. 1b) and could then be considered as the best representative of the full NMR set;[11] and model 13, which represents the opposite case with the largest average RMSD value in the experimental set. We also include the heat capacity curve resulting from the interaction model introduced in this manuscript (labeled as NMR-Global). We have focused this figure in the region around T_m, which would roughly correspond to the experimentally accessible range in a thermal denaturation experiment. In this figure, we have also included the symbols corresponding to the temperatures sampled along the parallel tempering process, and the error bars resulting from the averaging process of the individual runs for every case, to show the statistical accuracy of our simulation data. To complement the thermodynamic analysis of our transitions, we have used the WHAM method[33] to compute the free energy profiles at T_m as a function of the energy for the individual models, which are shown in Fig. S3 of the ESI.† In Fig. 3b, we show the results for the same selected cases shown in Fig. 3a. In this figure, we have normalized the energy scale dividing by the energy corresponding to the PDB conformer according to the energetic model employed in each case. This scale is similar to the fraction of native contacts Q usually employed in structure-based models, but does not depend on any arbitrary choice to determine when a native contact is present in a given sampled conformation.[34] This Q fraction has been established as an adequate reaction coordinate to study the folding process in structure-based models for proteins.[35] In our case, since the energy of a given native contact can continuously vary from its minimum value $(-\varepsilon)$ to zero, a formed contact can contribute with less than one reduced unit to the energy, and therefore the values we get for E/E_{nat} are somehow smaller than the standard values of Q usually reported. This is especially important at the transition temperature that corresponds to the free energy profiles shown in Fig. 3b, a relatively high temperature where fluctuations at the native state basin become important, especially for the models that explicitly treat these fluctuations.

It is interesting to mention that the X-ray structure provides the largest separation between the minima corresponding to the folded and unfolded states in equilibrium at the transition temperature, even at the reduced scale used in the x-axis of Fig. 3b. This is for sure a reflection of the more compact experimental structure, which partially reduces the structural fluctuations in the folded state even at a relatively high temperature as T_m. On the other hand, the results from individual NMR models are rather different among them (see Fig. S3 of the ESI† and Fig. 3b) and, depending on the conformer taken as reference, some of them could lead to the impression of an essentially downhill transition,[36] as NMR-13, or to transitions that at the temperature of the absolute maximum of the heat capacity correspond to an equilibrium between the unfolded and an intermediate state. In this latter case, the analysis of the trajectories (not shown) indicates that in this intermediate, at least the C-terminal tail of the protein is unfolded, and in some cases, the turn between β1 and α1 is highly distorted as well; they only become properly attached in their native positions at lower temperatures, mostly due to the effect of local interactions (which explains the second, shallow peak in the heat capacity curve appearing in some cases, as that of NMR-02 in Fig. 3a). It is interesting to mention that the results from the new model NMR-Global introduced in this work show a nice two-state transition with well-defined minima for the unfolded and folded states at T_m, as shown in Fig. 3b.

These differences in the folding landscape resulting from the distinct simulation models are better appreciated in Fig. 4, where two-dimensional landscapes are represented as a function of E/E_{nat} and the RMSD value between the sampled

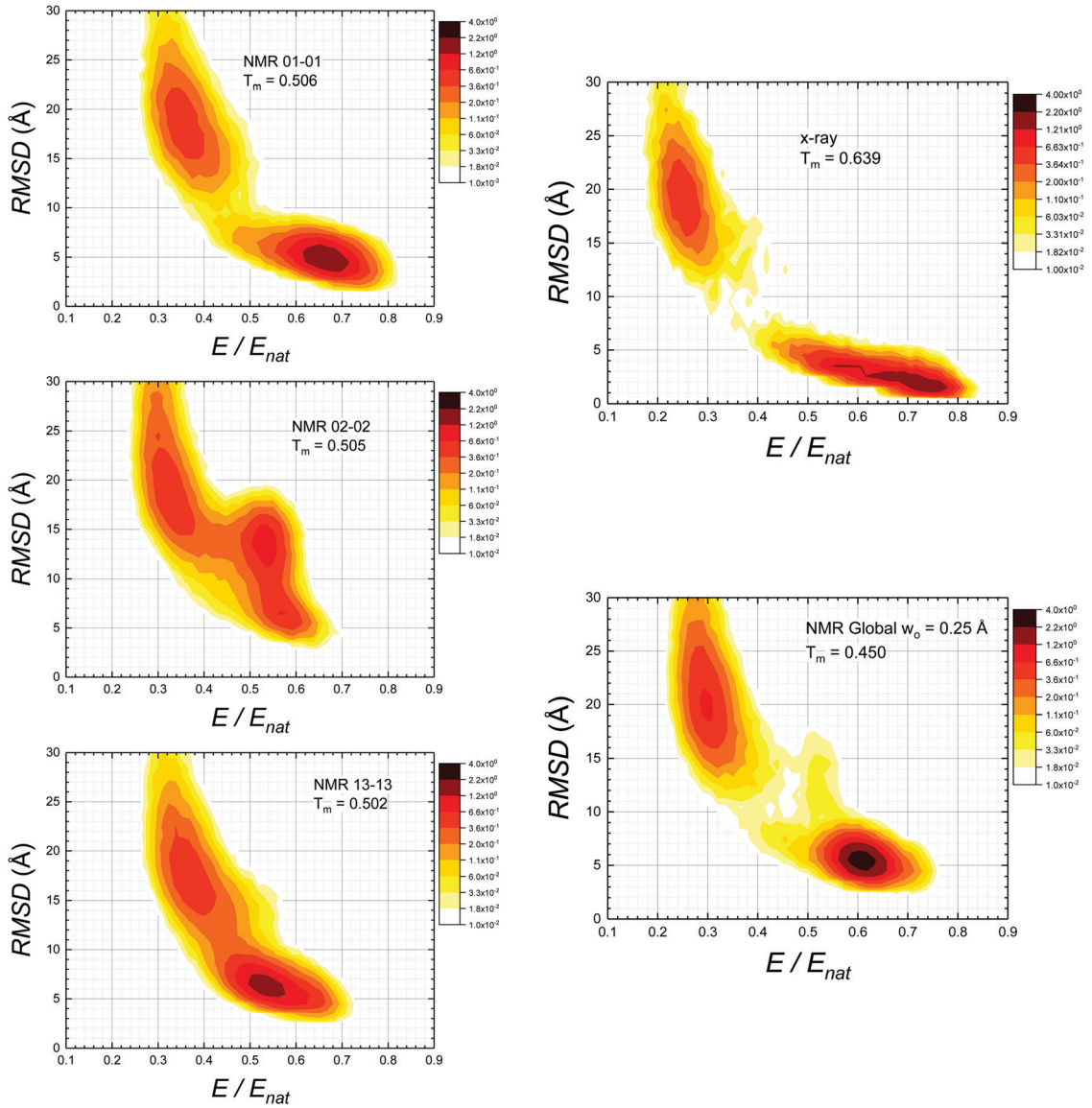


Fig. 4 Free energy landscapes, using as coordinates the reduced energy and the RMSD from the native conformations, obtained at the transition temperatures of the indicated models.

conformations and the reference experimental conformation (in NMR-Global, we have used the first NMR conformer as reference for this analysis). The X-ray results show a relatively wide minimum along the energy scale for the native state at T_m, which was also observed in other structure-based models considering the X-ray structure of protein S6.[12] In contrast, the NMR-Global model gives a well-defined minimum at its transition temperature, with fluctuations barely populated, although with slightly less ''nativeness'' than the X-ray case, as indicated by the position of the native minimum along the E/E_{nat} coordinate. The free energy barrier is not as large in the Global-NMR model as that resulting from the X-ray structure (which, as a matter of fact, is one of the largest barriers we have ever found using this simulation model[18,20]), but much larger than those provided by any of the individual NMR models. This is an interesting result, since it indicates that the NMR-Global model is not just a plain average of the individual NMR conformers, but instead creates new features, which, at least for this protein, provide results better than those arising from any of the individual conformations in the experimental NMR file.

As stated at the beginning of this section, we want to focus our analysis on the ability of the model to reproduce the conformational fluctuations found in the native state (at the level of the 20 NMR conformers in file 2KJV). To quantify the amplitude of these fluctuations resulting from our simulations along the protein sequence, we have chosen a lower temperature, in which the folded structure is the only one populated, and therefore the next results are computed at 90% of the corresponding T_m, which could be considered as an estimation of room temperature for this protein. We have used the root mean square fluctuation, RMSF, calculation included in VMD,[25]

using in each case the conformations sampled at the indicated temperature as a statistical ensemble, and the corresponding experimental conformer as a structural reference (again, for the NMR-Global model, the NMR-01 conformation is used). Just to check the methodology, in Fig. S4 of the ESI†, we show the individual values of the fluctuations after optimal superposition to the reference structure for several thousand conformations sampled at 0.9 T_m, equally scattered along one of the NMR-Global model trajectories. As can be seen, the results are quite consistent along the trajectory, and permit us to clearly detect the regions of large and small structural fluctuations.

The average results as a function of the sequence position are shown in Fig. 5, which is separated into two panels to favor the visualization of the results. Both panels show, in blue, the experimental RMSF corresponding to the superposition of the 20 NMR models in file 2KJV. It is then a quantitative reflection of the structural superposition shown in Fig. 1a, and as already mentioned, shows the largest experimental fluctuations in the C-terminal tail and in the middle loop. The X-ray model (dotted black line in the bottom panel) correctly shows a larger mobility in the C-terminal tail (partially precluded by the lack of several residues in the crystal structure in this region), but otherwise it presents very small fluctuations along the full sequence at this temperature. Even the maximum in the central loop is just marginal, and this indicates that the fact that this loop is bent towards the core of the structure in the X-ray conformation (see Fig. S1, ESI†), a feature which is absent in any of the NMR models, gives this region enough native contacts to stop it from fluctuating at room temperature; this fact does not seem to be correct in solution for this protein. The results from individual NMR models are, again, rather different from one another (top panel of Fig. 5). Any of them reproduces the large fluctuations in the C-terminal region, where, as can be seen in Fig. 2, there are no native contacts. However, the fluctuations of the middle loop have rather different amplitudes, and in some cases, depending on the exact position and number of native contacts for every model, large fluctuations appear in other regions, as happens in model NMR-02.

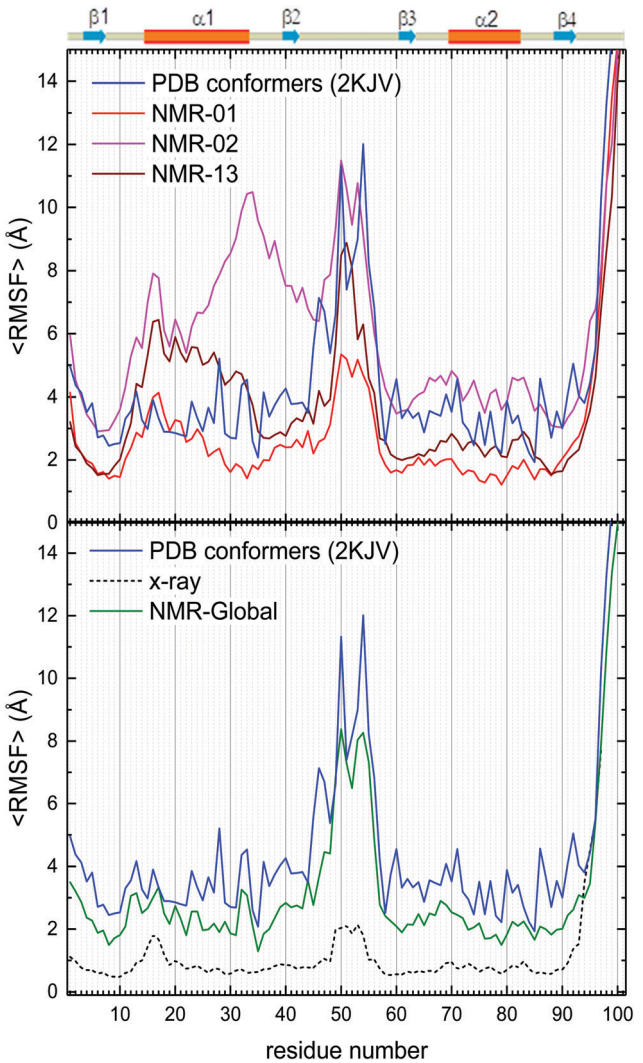


Fig. 5 Average values of the root mean square fluctuations with respect to native (RMSF, in Å) at the residue level, computed for the folded conformations of the different models (at T = 0.9 T_m).

On the contrary, the NMR-Global model (bottom panel) properly reproduces, at a semiquantitative level, the experimental fluctuations. A full quantitative coincidence is not reasonable, since it would depend on the exact temperature of the simulation results, as well as the experimental data and the conditions used in their interpretation. Of course, this good agreement for the fluctuations of the native state in the NMR-Global model is partly expected, since the experimental fluctuations are used to compute the structure-based force-field that the model uses. But, it is interesting to check that the methodological treatment used to include this information into the interaction potential, and to define from it a very simple and computationally efficient coarse-grained model, still preserves the experimental information used as input, and can therefore represent a proper methodology, in case better fluctuation information was available.

Conclusions

In this manuscript, we have introduced a new structure-based simulation model for protein folding. Its main novelty lies in the fact that it uses several conformers for the native state, therefore allowing us to consider protein flexibility as an inherent part of the definition of the interaction model. As a test, we have considered the different conformations appearing in the NMR-solved structure of S6 ribosomal protein.

NMR conformers, as collected in PDB files, have been recently used as a signature of fluctuations in the folded state of proteins in a series of physics-based analyses of the native state.[37] However, they may not be a good reflection of equilibrium populations of conformers in the native state at room temperature. Probably, physically sounder ensembles could be obtained from equilibrium molecular simulations, and several procedures and databases are already available for that aim.[38–40] Nevertheless, at least for the S6 protein used as a test

in this work, they provide rather static structures, with structural fluctuations that are very small in comparison with the experimental flexibility provided by the NMR file 2KJV, and whose folding results are then not so different from those provided by traditional single structure approaches. This fact could be related to the relatively short simulation time used by the methods mentioned, which are not able to sample large conformational fluctuations. Since the determination of ''contact distances'' for the native contacts is, in our experience, a critical part of the definition of the interaction model based on flexible structures, we have preferred to keep the experimental information of the NMR PDB file as representative.

The model uses a coarse-grained, α-carbon representation of the polypeptide chain. A Monte Carlo method, coupled to a parallel tempering procedure, is used to analyze the behavior of the system and the folding/unfolding transition of the protein as a function of temperature.

The results for the new model, termed NMR-Global here, provide evidence for a two-state folding transition for this protein, with a relatively high free energy barrier, in agreement with experimental evidence.[28] The results are, in this sense, much better than those provided by standard models that use a single NMR conformer to define the contact map and, therefore, the interaction potential for the simulation. In these cases, negligible barriers appear in several cases, and poorly defined folded states are populated at the transition temperature.

The fluctuations found for the native state of the new simulation model at room temperature also reproduce the regions of the sequence that are more flexible in the experimental data used to define the interactions of NMR-Global. Again, the results of the new model are, in this sense, much better than those coming from single structure NMR models, or from the X-ray single conformation.

Therefore, we consider that the model is ready for further exploration, and its use in future work in many more flexible protein structures will provide the final test of its applicability to the study of the protein folding process, always at the coarse-grained level of a structure-based model. Although the experimental structures used to define the fluctuations of the protein are only approximate if taken from an NMR PDB file, the methodology introduced in this manuscript can be readily applied to any set of conformations that are representative of the structural flexibility of the native state. At the coarse-grained level used in the model, the definition of the global contact map and the distance analysis of the native contacts are very fast for several tens of conformations, a standard number in NMR PDB files,[6] and can be easily automated so that a larger number is considered if available.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

We acknowledge financial support from ''Proyectos de Investigación Santander-UCM'' (PR26/16-20251) and from Spanish ''Ministerio de Economía y Competitividad'' (under grant CTQ2016-78895-R).

References

1 H. Taketomi, Y. Ueda and N. Gō, *Int. J. Pept. Protein Res.*, 2009, **7**, 445–459.
2 R. D. Hills, Jr. and C. L. Brooks, 3rd, *Int. J. Mol. Sci.*, 2009, **10**, 889–905.
3 S. Kmiecik, D. Gront, M. Kolinski, L. Wieteska, A. E. Dawid and A. Kolinski, *Chem. Rev.*, 2016, **116**, 7898–7936.
4 S. Riniker, J. R. Allison and W. F. van Gunsteren, *Phys. Chem. Chem. Phys.*, 2012, **14**, 12423–12430.
5 S. G. Estacio, H. Krobath, D. Vila-Vicosa, M. Machuqueiro, E. I. Shakhnovich and P. F. Faisca, *PLoS Comput. Biol.*, 2014, **10**, e1003606.
6 H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov and P. E. Bourne, *Nucleic Acids Res.*, 2000, **28**, 235–242.
7 G. S. Rule and T. K. Hitchens, *Fundamentals of Protein NMR Spectroscopy*, Springer, 2006.
8 R. Sharma, D. De Sancho and V. Munoz, *Phys. Chem. Chem. Phys.*, 2017, **19**, 28512–28516.
9 B. Li, M. Fooksa, S. Heinze and J. Meiler, *Crit. Rev. Biochem. Mol. Biol.*, 2018, **53**, 1–28.
10 M. F. Rey-Stolle, M. Enciso and A. Rey, *J. Comput. Chem.*, 2009, **30**, 1212–1219.
11 P. Jiang and U. H. Hansmann, *J. Chem. Theory Comput.*, 2012, **8**, 2127–2133.
12 H. Lammert, J. K. Noel, E. Haglund, A. Schug and J. N. Onuchic, *J. Chem. Phys.*, 2015, **143**, 243141.
13 R. B. Best, Y. G. Chen and G. Hummer, *Structure*, 2005, **13**, 1755–1763.
14 K. Okazaki, N. Koga, S. Takada, J. N. Onuchic and P. G. Wolynes, *Proc. Natl. Acad. Sci. U. S. A.*, 2006, **103**, 11844–11849.
15 C. D. Bope, D. Tong, X. Li and L. Lu, *Prog. Biophys. Mol. Biol.*, 2017, **128**, 100–112.
16 P. Weinkam, J. Pons and A. Sali, *Proc. Natl. Acad. Sci. U. S. A.*, 2012, **109**, 4875–4880.
17 L. Prieto, D. de Sancho and A. Rey, *J. Chem. Phys.*, 2005, **123**, 154903.
18 L. Prieto and A. Rey, *J. Chem. Phys.*, 2007, **127**, 175101.
19 Y. Sugita and Y. Okamoto, *Chem. Phys. Lett.*, 1999, **314**, 141–151.
20 M. Larriva, L. Prieto, P. Bruscolini and A. Rey, *Proteins*, 2010, **78**, 73–82.
21 M. Enciso and A. Rey, *Biophys. J.*, 2011, **101**, 1474–1482.
22 M. A. Soler, A. Rey and P. F. Faisca, *Phys. Chem. Chem. Phys.*, 2016, **18**, 26391–26403.
23 A. Ohman, T. Oman and M. Oliveberg, *Protein Sci.*, 2010, **19**, 183–189.
24 M. Lindahl, L. A. Svensson, A. Liljas, S. E. Sedelnikova, I. A. Eliseikina, N. P. Fomenkova, N. Nevskaya, S. V. Nikonov,

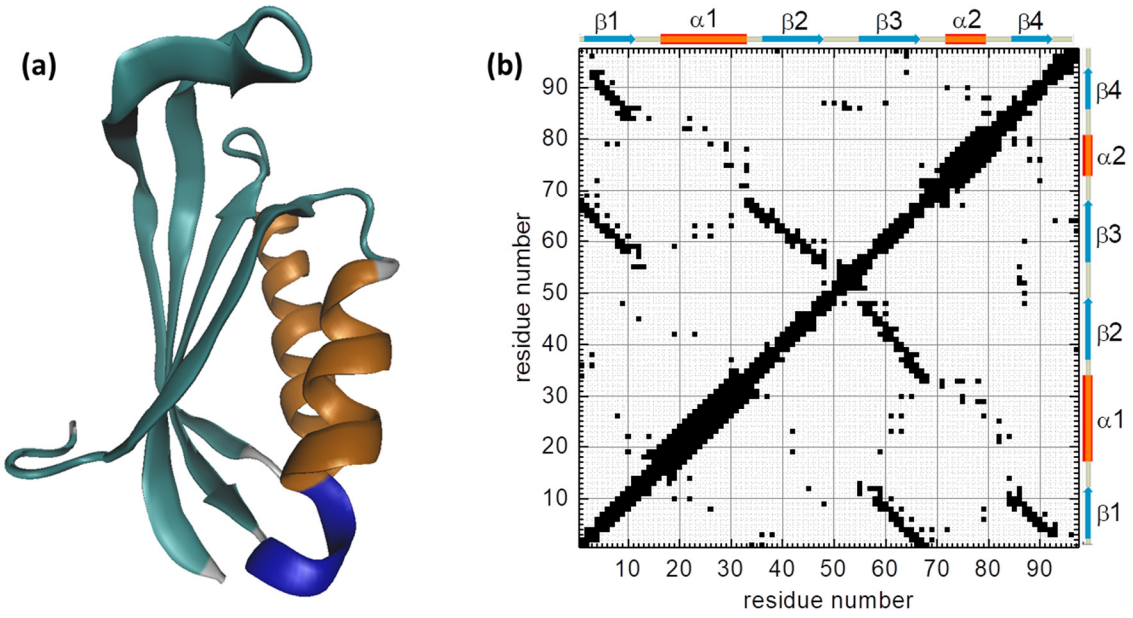
M. B. Garber and T. A. Muranova, *et al.*, *EMBO J.*, 1994, **13**, 1249–1254.
25 W. Humphrey, A. Dalke and K. Schulten, *J. Mol. Graphics*, 1996, **14**, 33–38.
26 A. Perez, A. Roy, K. Kasavajhala, A. Wagaman, K. A. Dill and J. L. MacCallum, *Proteins*, 2014, **82**, 2671–2680.
27 L. Prieto and A. Rey, *J. Chem. Phys.*, 2007, **126**, 165103.
28 D. E. Otzen, O. Kristensen, M. Proctor and M. Oliveberg, *Biochemistry*, 1999, **38**, 6499–6511.
29 M. O. Lindberg, J. Tangrot, D. E. Otzen, D. A. Dolgikh, A. V. Finkelstein and M. Oliveberg, *J. Mol. Biol.*, 2001, **314**, 891–900.
30 M. O. Lindberg and M. Oliveberg, *Curr. Opin. Struct. Biol.*, 2007, **17**, 21–29.
31 I. A. Hubner, M. Oliveberg and E. I. Shakhnovich, *Proc. Natl. Acad. Sci. U. S. A.*, 2004, **101**, 8354–8359.
32 M. Iwakura, T. Nakamura, C. Yamane and K. Maki, *Nat. Struct. Biol.*, 2000, **7**, 580.
33 J. D. Chodera, W. C. Swope, J. W. Pitera, C. Seok and K. A. Dill, *J. Chem. Theory Comput.*, 2007, **3**, 26–41.
34 K. Wolek and M. Cieplak, *J. Chem. Phys.*, 2016, **144**, 185102.
35 R. B. Best, G. Hummer and W. A. Eaton, *Proc. Natl. Acad. Sci. U. S. A.*, 2013, **110**, 17874–17879.
36 M. M. Garcia-Mira, M. Sadqi, N. Fischer, J. M. Sanchez-Ruiz and V. Munoz, *Science*, 2002, **298**, 2191–2195.
37 Q. Y. Tang, Y. Y. Zhang, J. Wang, W. Wang and D. R. Chialvo, *Phys. Rev. Lett.*, 2017, **118**, 088102.
38 M. Rueda, C. Ferrer-Costa, T. Meyer, A. Perez, J. Camps, A. Hospital, J. L. Gelpi and M. Orozco, *Proc. Natl. Acad. Sci. U. S. A.*, 2007, **104**, 796–801.
39 J. Camps, O. Carrillo, A. Emperador, L. Orellana, A. Hospital, M. Rueda, D. Cicin-Sain, M. D'Abramo, J. L. Gelpi and M. Orozco, *Bioinformatics*, 2009, **25**, 1709–1710.
40 M. Jamroz, A. Kolinski and S. Kmiecik, *Nucleic Acids Res.*, 2013, **41**, W427–W431.

Electronic Supplementary Material (ESI) for Physical Chemistry Chemical Physics.

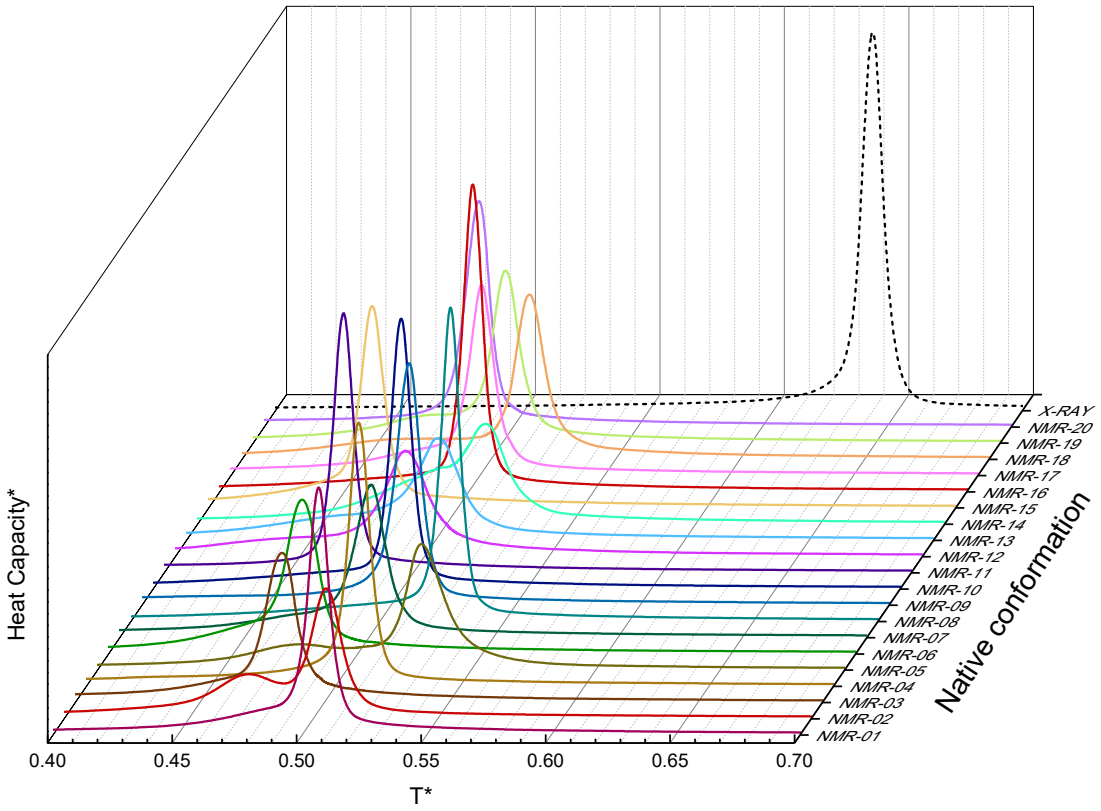
Design of a structure-based model for protein folding from flexible conformations

Ana M. Rubio and Antonio Rey*
Departamento de Química Física, Facultad de Ciencias Químicas, Universidad Complutense de Madrid, E-28040 Madrid, Spain.

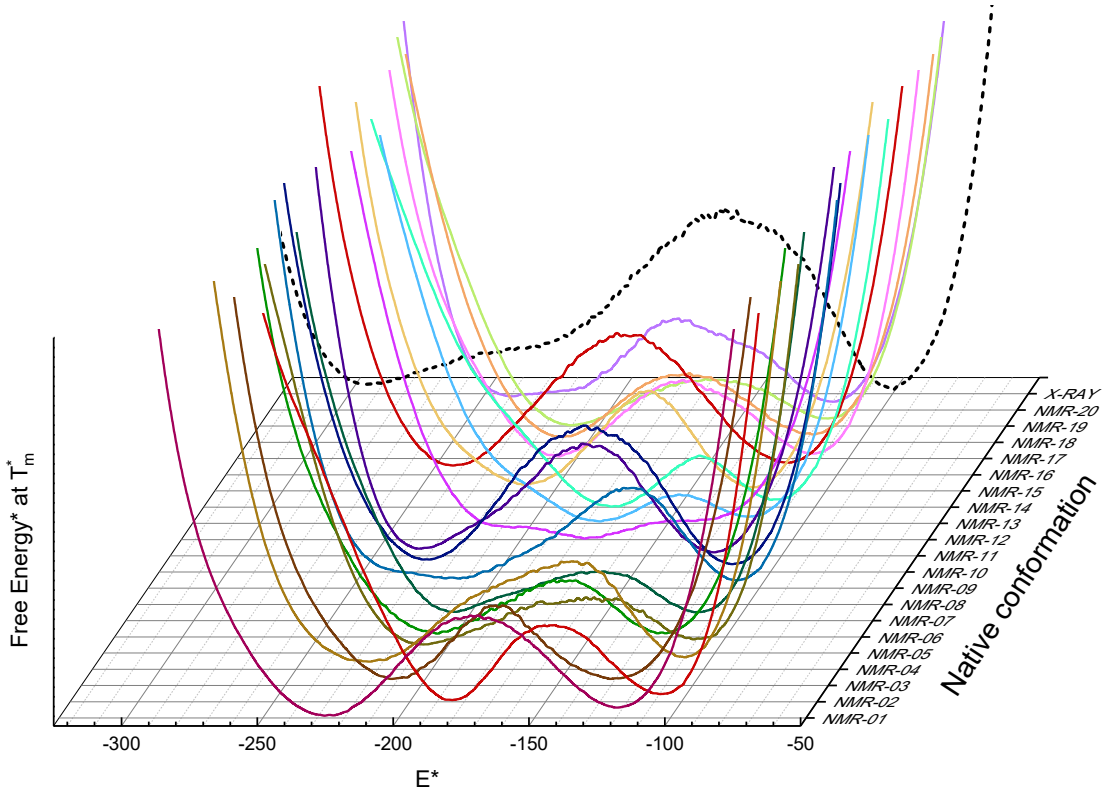
ELECTRONIC SUPPLEMENTARY INFORMATION (ESI)



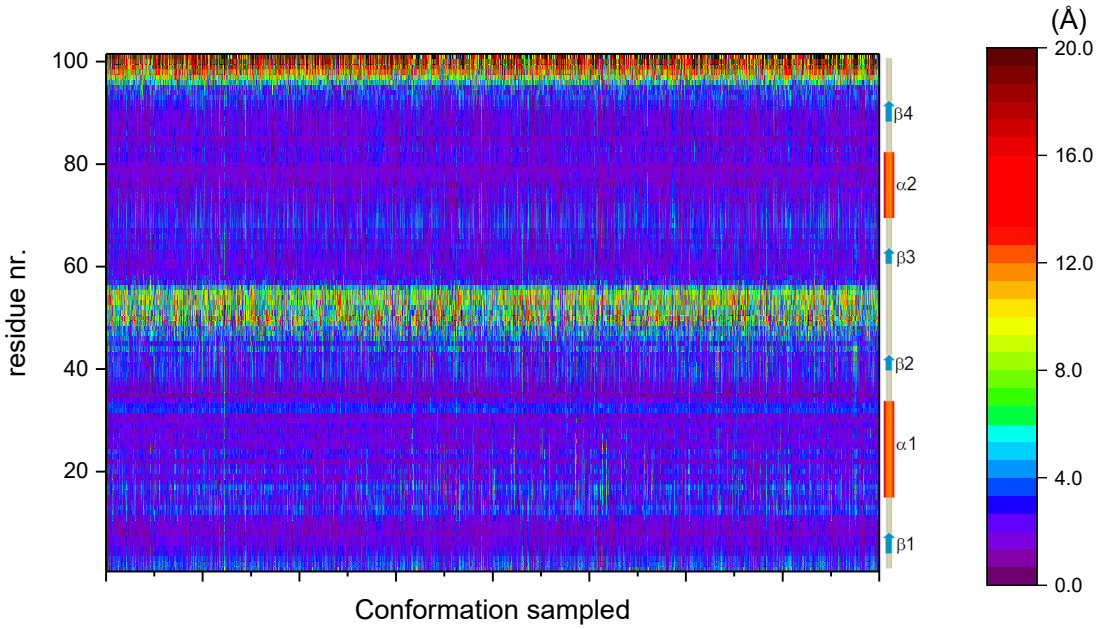
Supplementary Figure S1. Ribbon diagram and contact map for the x-ray structure of protein S6, as taken from the PDB file 1RIS. The positions of the secondary structure elements along the sequence are sketched at the axes of the contact map.



Supplementary Figure S2. Heat capacity curves (in reduced units) as a function of the reduced temperature, from standard structure-based calculations (based on a single structure, as indicated in the right axis).



Supplementary Figure S3. Free energy curves (in reduced units) as a function of reduced energy, from standard structure-based calculations (based on a single structure, as indicated in the right axis). The native basin appears at the left side (lower energy).



Supplementary Figure S4. Plot of the structural fluctuations after optimal superposition for a representative set of 8000 conformations along the simulation for T = 0.9 T_m in the NMR-Global model. The NMR-01 conformer from the PDB file 2KJV is used as reference structure.